COMP559 Homework 2 Solutions

Syed Saqib Habeeb
SH212

February 21, 2025

1 Problem 1

1.1 The first and second eigenvectors of the laplacian provide a good visualization of the data

No, the first and second eigenvectors of the laplacian do not provide a good visualization of the data. The first eigenvalue of a laplacian is always zero which means that the first eigenvector is constant and it does not provide us with any information. We can use the second and third eigenvectors for visualization.

1.2 The number of common neighbors between pairs of vertices is a good heuristic for link prediction in a BA graph

No, since the BA graphs have low clustering because of its preferential attachment. The common neighbour heuristic is generally a good heuristic for dense graphs. The common neighbour heuristic can still be used to identify some dense subgraphs but it will not give a good output for the full graph.

1.3 Isomap and Laplacian Eigenmaps are inductive embedding methods

No, these two are transductive embedding methods. They work by leveraging the structure of the specific dataset through graph-based techniques and eigenvalue or distance computations and they don't produce a generalizable mapping function.

1.4 Deep learning has solved the curse of dimensionality

No, while deep learning has reduced the curse of dimesionality universally, it has not solved it completely. The curse of dimesionality still very much exists when the data is unstructured and consists of high dimensional noise.

1.5 Eigenvalues of the Laplacian can be used to compare graphs in terms of topological similarity even if they have different sizes

Yes, from the concepts of spectral graph theory we can say that eigenvalues of the laplacian can be used to compare graphs in terms of topological similarity even if they have different sizes. The number of eigenvalues will differ between graphs of different sizes. However, we can still compare the overall spectral distributions or focus on specific eigenvalues.

2 Problem 3

I performed comparisions on both unscaled and scaled datasets (I scaled the datasets using `MinMaxScaler()`). Refer the code here. Figures 1 and 3 are comparisions on the original datasets and figures 2 and 4 are comparisions of the different methods after scaling the datasets using `MinMaxScaler()`.
For the plots, the hyperparameters set for each method were similar and as follows

- For Isomap *n_components* was set to 2 and *n_neighbors* was set to 10
- For Laplacian *n_components* was set to 2 and *n_neighbors* was set to 10
- For TSNE *n_components* was set to 2 and *perplexity* was set to 10
- For Autoencoder *encoding_dim* was set to 2, *batch_size* was set to 32 and *epochs* was set to 50
- I also compared the performance when we change the value of *n_neighbors*

As we can see from figures 1, 2, 3 and 4 we can conclude the following from the plots obtained about the performance of each method.
Isomap:

- **For Swiss roll:**
  - **Advantages**
    - Effectively unrolls the manifold by capturing the global geometric structure
  - **Disadvantages:**
    - Sensitive to the choice of the number of neighbors
- **For Cylinder:**
  - **Advantages**
    - Can capture the circular structure preserving the continuity of the cylindrical surface
  - **Disadvantages:**
    - If the graph construction is off (e.g., inappropriate neighborhood size), it might not fully capture the circularity, leading to distortions in the embedding
- **For Digits:**
  - **Advantages**
    - Can reveal global relationships among digits
  - **Disadvantages:**
    - The high intrinsic variability and overlapping classes in digits can lead to embeddings that do not clearly separate clusters

Laplacian:

- **For Swiss roll:**
  - **Advantages**
    - Excellent at maintaining local geometric relationships, which helps to capture the manifold's local curvature
  - **Disadvantages:**
    - Tends to lose global information, which might make the overall unfolding of the Swiss roll less apparent
- **For Cylinder:**
  - **Advantages**
    - Captures local circular neighborhoods effectively
  - **Disadvantages:**
    - The overall global structure (the continuity of the cylinder) may not be as clearly preserved because the focus is on local rather than global distances
- **For Digits:**
  - **Advantages**
    - Succeeds in grouping together digits that share similar local features
  - **Disadvantages:**
    - Global separation of distinct digit clusters may be less pronounced, making it sometimes harder to visually distinguish between classes

TSNE:

- **For Swiss roll:**
  - **Advantages**
    - Can reveal fine-grained local clusters within the Swiss roll, showing clear groupings of similar regions
  - **Disadvantages:**
    - Distorts the global geometry
- **For Cylinder:**
  - **Advantages**
    - Excels at delineating small local neighborhoods
  - **Disadvantages:**
    - May artificially separate continuous structures, leading to an embedding that appears to fragment the cylinder into multiple parts
- **For Digits:**

- **Advantages**
  * Creates well-separated clusters
- **Disadvantages:**
  * Computationally intensive, especially with larger datasets

Autoencoder:

- **For Swiss roll:**
  - **Advantages**
    * Can learn non-linear transformations and, if well-tuned, capture both local and global structures
  - **Disadvantages:**
    * Requires careful tuning of architecture, hyperparameters, and training duration
- **For Cylinder:**
  - **Advantages**
    * The autoencoder can capture periodicity and the overall cylindrical structure
  - **Disadvantages:**
    * Training instabilities or convergence issues might arise
- **For Digits:**
  - **Advantages**
    * Provides a powerful way to learn latent representations that capture intricate variations in handwriting
  - **Disadvantages:**
    * Risk of learning representations that focus on reconstruction rather than on clear cluster separation

Therefore, we can conclude that

- Isomaps perform best when have to preserve the global geometry but its reliance on the neighborhood graph makes it sensitive to parameter choices
- Laplacians can be used when we have to capture small-scale structures but fails to capture any global structure
- TSNE is exceptional for revealing local clusters, especially in high-dimensional data however it sacrifices global distance fidelity and can produce different visualizations based on parameter settings
- Autoencoders provide a flexible, learning-based approach that can capture both local and global structure if properly designed and trained

PS: All the plots plotted are available here.

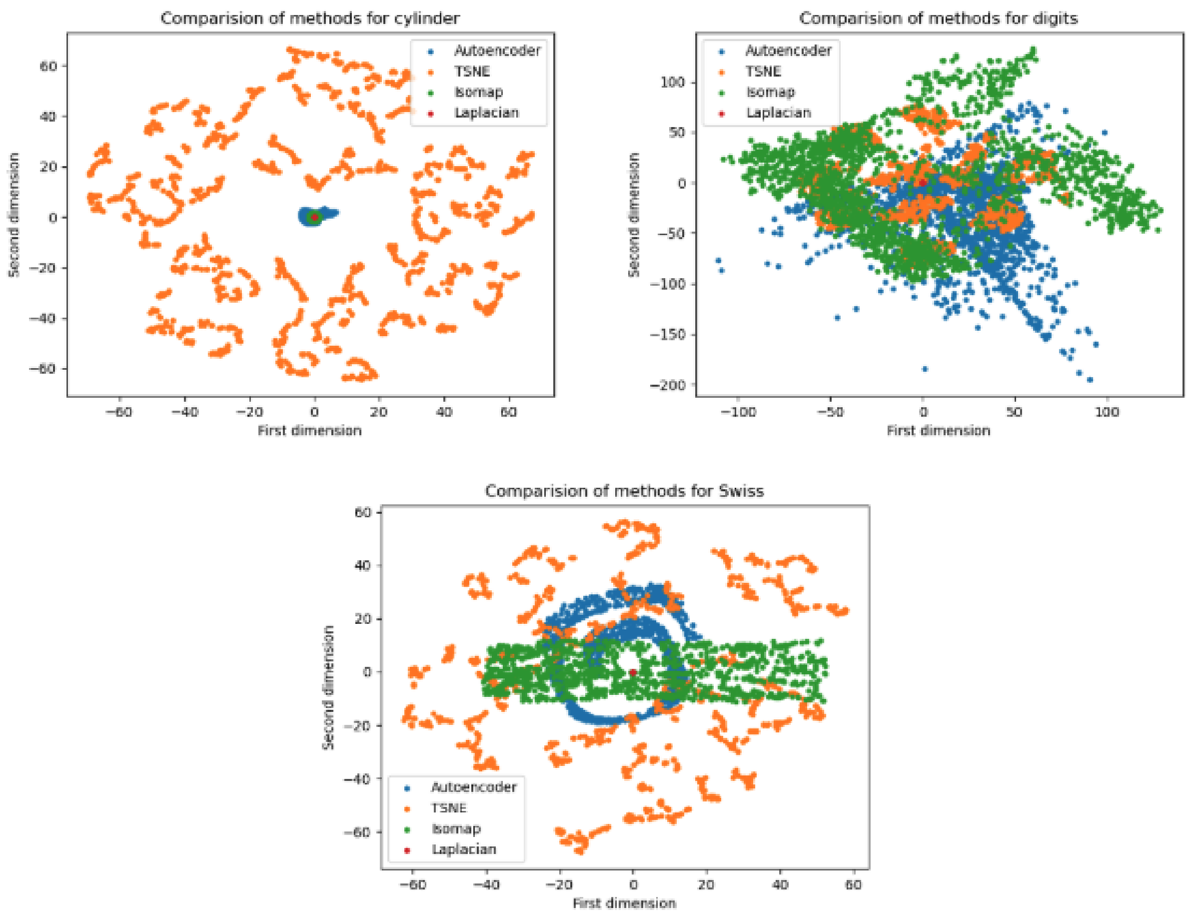

Figure 1: Comparision of different methods

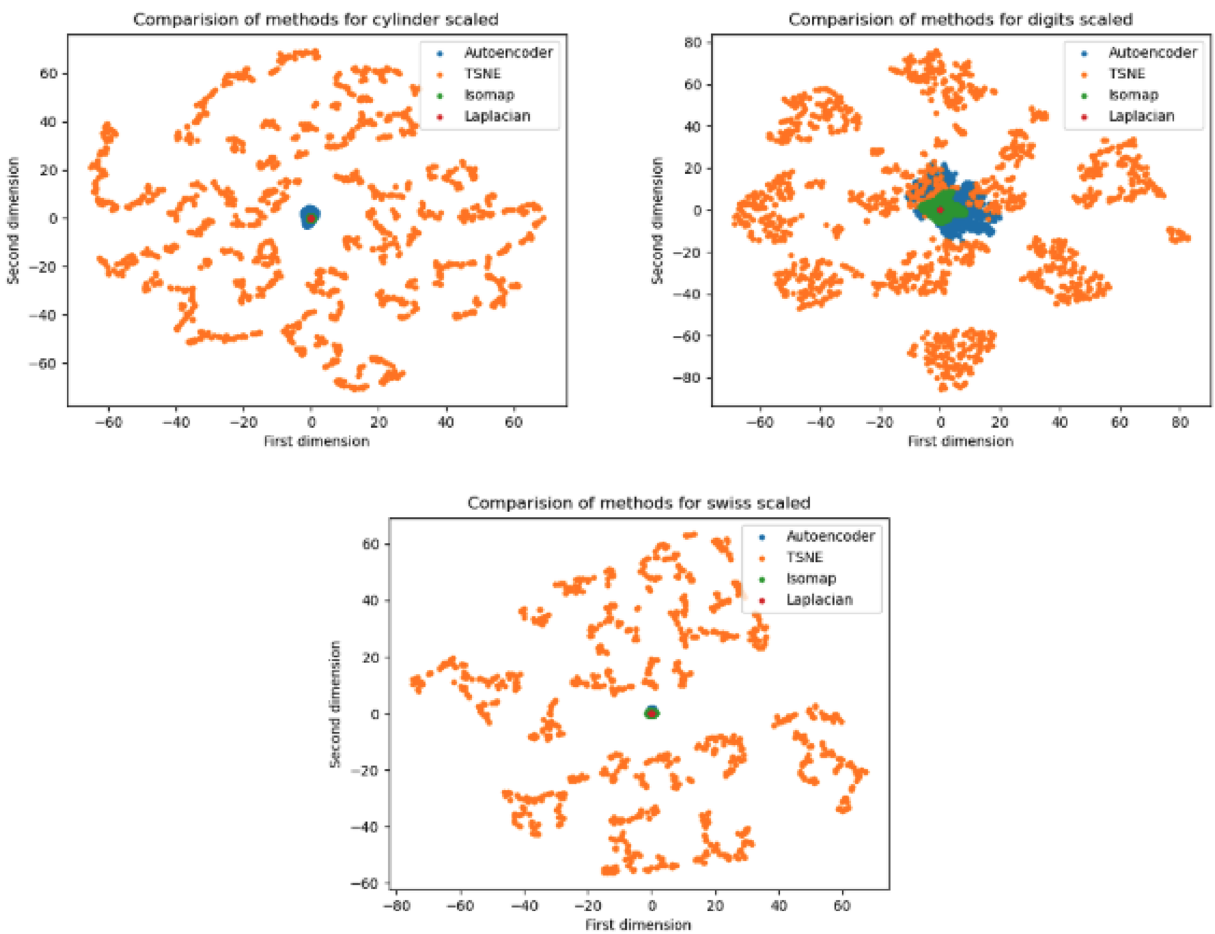

Figure 2: Comparision of different methods on scaled datasets

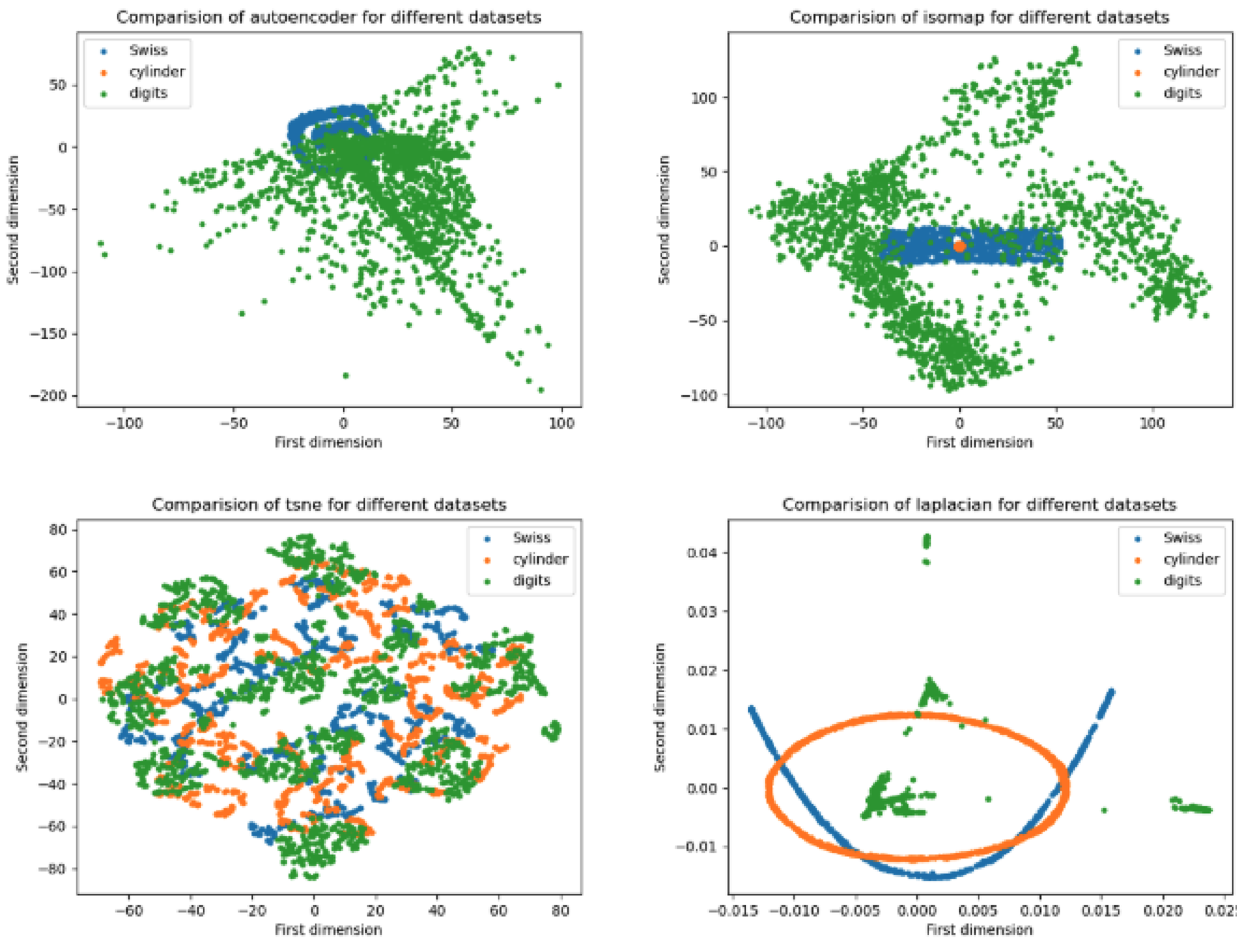

Figure 3: Comparision of the same method on different datasets

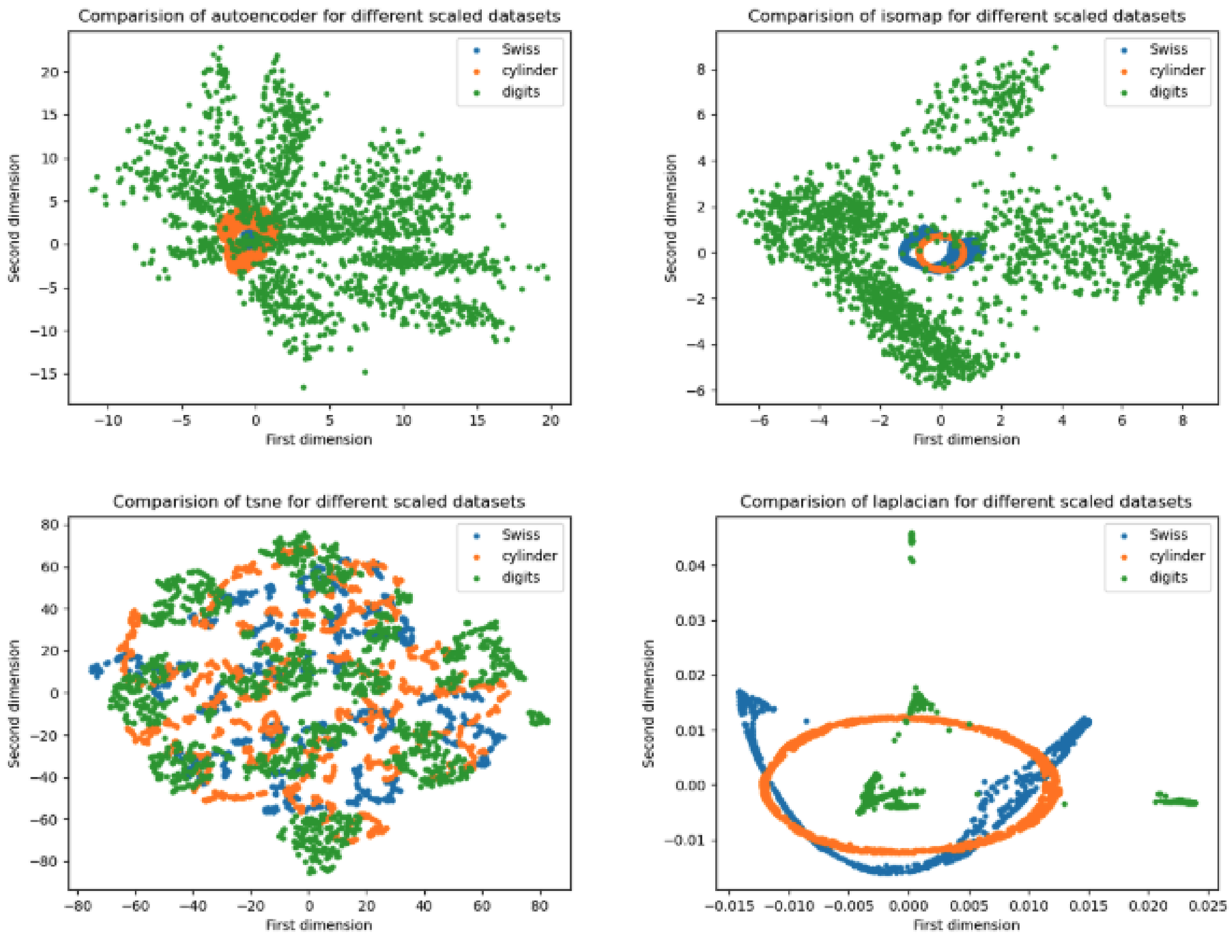

Figure 4: Comparision of the same method on different scaled datasets